PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-273214
(43)Date of publication of application : 05.10.2001

(51)Int.Cl. G06F 13/00
G06F 17/21

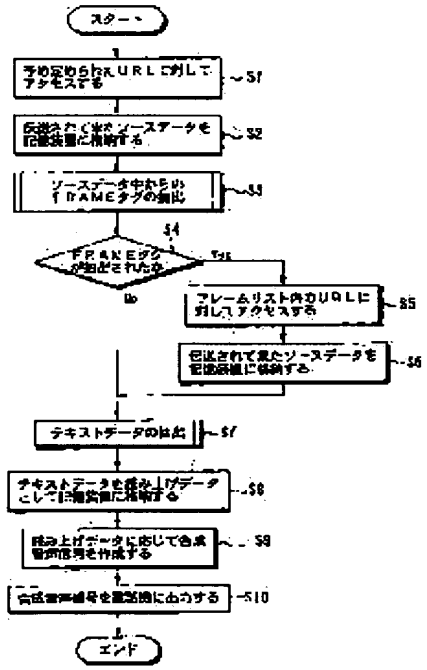
(21)Application number : 2000-083150
(22)Date of filing : 24.03.2000

(71)Applicant : OKI SOFTWARE KK
OKI ELECTRIC IND CO LTD
NTT ME CORP

(72)Inventor : NISHIMURA HITOSHI
YAKIDA KAZUHIKO
MATSUSHITA ARIYUKI
YAMAGUCHI YUICHIRO
ITO SHINICHI
NAGAI TOMOYASU
OTA TSUYOSHI
TAMURA MASARU

(54) WEB PAGE DECODING SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a Web page decoding system capable of exactly extracting a text part in respect to the document of an HTML containing a tag having the description of a uniform resource locator(URL).
SOLUTION: Basic source data comprising a Web page are extracted from a storage area designated by the prescribed URL, and written in a storage means and when the existence of the tag containing the description of the URL is detected out of the basic source data, the URL in that prescribed tag is detected. Then, source data are extracted from the storage area designated by that detected URL, and written in the storage means and the text part is extracted from all the source data stored in the storage means.



(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－273214
（P2001－273214A）
(43)公開日　平成13年10月5日(2001.10.5)

(51)Int.Cl.⁷	識別記号	FI		テーマコード(参考)
G06F 13/00	354	G06F 13/00	354D	5B009
17/21	501	17/21	501T	5B089
	568		568A	
	596		596A	

審査請求　未請求　請求項の数6　ＯＬ　（全 6 頁）

(21)出願番号　特願2000－83150(P2000－83150)

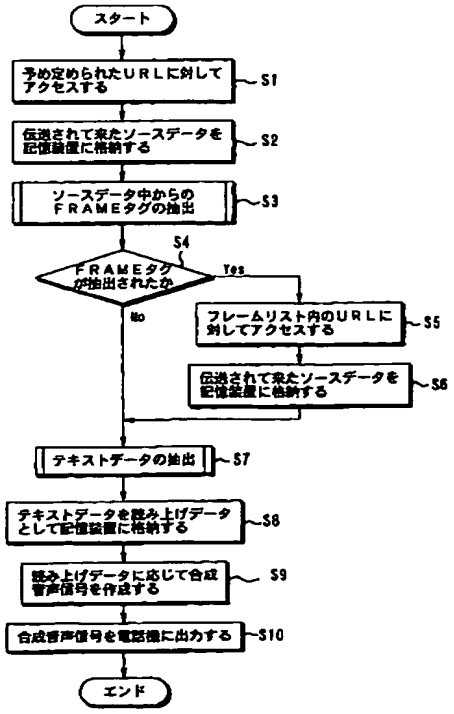
(22)出願日　平成12年３月24日(2000.3.24)

(71)出願人　591051645
沖ソフトウェア株式会社
東京都板橋区舟渡１丁目12番11号
(71)出願人　000000295
沖電気工業株式会社
東京都港区虎ノ門１丁目７番12号
(71)出願人　596094692
株式会社エヌ・ティ・ティ　エムイー
東京都千代田区大手町二丁目２番２号
(74)代理人　100079119
弁理士　藤村　元彦

最終頁に続く

(54)【発明の名称】　ウェブページ解読システム

(57)【要約】
【課題】　ＵＲＬ(ユニホームリソースロケータ)の記述を有するタグを含むＨＴＭＬの文書に対してテキスト部分を正確に抽出することができるＷｅｂ（ウェブ）ページ解読システムを提供する。
【解決手段】　Ｗｅｂページを構成する基本のソースデータを所定のＵＲＬで指定された記憶領域から取り出して記憶手段に書き込み、基本のソースデータ中からＵＲＬの記述箇所を含む所定のタグの存在を検出した場合には、その所定のタグ中のＵＲＬを検出し、その検出したＵＲＬで指定された記憶領域からソースデータを取り出して記憶手段に書き込み、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出する。



1

【特許請求の範囲】

【請求項１】　Ｗｅｂ（ウェブ）ページを構成するＨＴＭＬ文書のテキスト部分を解読するＷｅｂページ解読システムであって、

前記Ｗｅｂページを構成する基本のソースデータを所定のＵＲＬ(ユニホームリソースロケータ)で指定された記憶領域から取り出して記憶手段に書き込む手段と、

前記基本のソースデータ中からＵＲＬの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、

前記所定のタグの存在が検出された場合にはその所定のタグ中のＵＲＬを検出するＵＲＬ検出手段と、

前記ＵＲＬ検出手段によって検出されたＵＲＬで指定された記憶領域からソースデータを取り出して前記記憶手段に書き込む手段と、

前記記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴とするウェブページ解読システム。

【請求項２】　前記所定のタグはフレームタグであることを特徴とする請求項１記載のウェブページ解読システム。

【請求項３】　前記ＵＲＬ検出手段は、前記フレームタグ中のＵＲＬとして＜ＦＲＡＭＥ　ＳＲＣ＝"ＵＲＬ"＞の構文中のＵＲＬを検出することを特徴とする請求項１又は２記載のウェブページ解読システム。

【請求項４】　前記テキスト抽出手段は、前記記憶手段に記憶されたソースデータ中の＜＞で囲まれた部分以外の部分をテキスト部分として抽出することを特徴とする請求項１記載のウェブページ解読システム。

【請求項５】　前記テキスト抽出手段によって抽出されたテキスト部分に対応して音声信号を作成して出力する音声出力手段を更に備えたことを特徴とする請求項１記載のウェブページ解読システム。

【請求項６】　前記音声出力手段の出力音声信号は公衆電話回線を介して電話機に供給されることを特徴とする請求項１記載のウェブページ解読システム。

【発明の詳細な説明】

【０００１】

【発明が属する技術分野】本発明は、Ｗｅｂ（ウェブ）ページのテキスト部分を解読するＷｅｂページ解読システムに関する。

【０００２】

【従来の技術】インターネットの情報サービスの１つであるWWW(World Wide Web)は、ＨＴＭＬ(Hyper Text Markup Language)という言語で記述されたＨＴＭＬファイルとそのファイルの保存位置の識別子であるＵＲＬ(Uniform Resource Locator)とを用いてインターネットを介して文字、映像、音声等のマルチメディア情報を参照することができるものである。ＨＴＭＬファイルをＷＷＷブラウザと呼ばれる閲覧ソフトウエアによって処理することによりディスプレイ画面上に形成されるものがＷｅｂページである。WWWの情報を提供する側、すなわちWWWサーバはＨＴＭＬファイルをＵＲＬで関連付けて保存しており、サーバアプリケーションに従って動作する。情報を提供される側、すなわちクライアント（クライアントコンピュータ）ではWWWブラウザを用いて所望のＵＲＬからＨＴＭＬファイルを含むソースデータ（例えば、画像ファイルや音声ファイル）にインターネットを介してアクセスしてソース中のファイルによって形成されたＷｅｂページをディスプレイ画面上で参照することができる。

2

【０００３】ＨＴＭＬの文書は、通常、テキスト文書からなるテキスト部とタグによって形成される表示情報部とから形成される。タグは＜＞を一対とする記号であり、タグを用いてＨＴＭＬの構文が例えば、＜ＨＴＭＬ＞～＜／ＨＴＭＬ＞の如く形成される。タグ＜＞で囲まれた部分にはＷｅｂページに表示されるテキスト部の文字の大きさ、フォントの種類、その文字色、Ｗｅｂページの背景色、画像ファイル名、画像位置等の様々な表示情報が示される。

【０００４】このようにＨＴＭＬの文書においてはタグ＜＞で囲まれた部分は表示される部分ではなく、文書を表示するための制御情報であるので、タグ＜＞で囲まれた部分を除くと、通常、単なるテキスト文書となることが普通である。一方、WWWブラウザに表示されるＷｅｂページの文書を読み上げるシステムがインターネット上には形成されることがある。これは、クライアントの端末がコンピュータではなく、例えば、公衆回線に接続された電話機である場合にＷｅｂページの文書を読み上げて音声信号として電話機に送出するためである。Ｗｅｂページ読上としては読み上げ対象のＨＴＭＬファイル中からタグ＜＞で囲まれた部分を除くテキストデータ部分を抽出し、そのテキストデータ部分の文字コードに対応した音声データを合成して一連の音声信号として出力することが行われる。

【０００５】

【発明が解決しようとする課題】ＨＴＭＬには、Ｗｅｂページ上にフレームを形成するための構文として、例えば、＜ＦＲＡＭＥＳＥＴ＞～＜／ＦＲＡＭＥＳＥＴ＞があり、これを用いたＷｅｂページでは分割された画面が得られる。その構文が記述されたＨＴＭＬファイルからは分割画面毎に別のＨＴＭＬファイルが更に呼び出されて文書が表示される。すなわち、＜ＦＲＡＭＥ　ＳＲＣ＝"ＵＲＬ"＞の如きタグにより分割画面毎にＵＲＬ（ＨＴＭＬファイル名を含む）が指定され、その指定されたＵＲＬの領域に存在するＨＴＭＬファイルの内容が表示される。

【０００６】しかしながら、従来、このようなフレームタグのようにＵＲＬの記述を有するタグを含むＨＴＭＬの文書に対してはテキスト部分を正確に抽出することができないという問題点があった。そこで、本発明の目的

は、ＵＲＬの記述を有するタグを含むＨＴＭＬの文書に対してテキスト部分を正確に抽出することができるＷｅｂページ解読システムを提供することである。

【０００７】

【課題を解決するための手段】本発明のＷｅｂページ解読システムは、Ｗｅｂページを構成するＨＴＭＬ文書のテキスト部分を解読するＷｅｂページ解読システムであって、Ｗｅｂページを構成する基本のソースデータを所定のＵＲＬ(ユニホームリソースロケータ)で指定された記憶領域から取り出して記憶手段に書き込む手段と、基本のソースデータ中からＵＲＬの記述箇所を含む所定のタグの存在を検出するタグ検出手段と、所定のタグの存在が検出された場合にはその所定のタグ中のＵＲＬを検出するＵＲＬ検出手段と、ＵＲＬ検出手段によって検出されたＵＲＬで指定された記憶領域からソースデータを取り出して記憶手段に書き込む手段と、記憶手段に記憶されたソースデータの全てからテキスト部分を抽出するテキスト抽出手段と、を備えたことを特徴としている。この構成より、基本のソースデータがフレームタグを含むＨＴＭＬファイルの場合には、そのフレームタグ内に記述されたＵＲＬの領域に格納されているＨＴＭＬファイルのテキスト部分も抽出することができる。

【０００８】

【発明の実施の形態】以下、本発明の実施例を図面を参照しつつ詳細に説明する。図１は本発明のによるＷｅｂページ解読システムの構成を示している。このシステムにおいては、WWWサーバ１は情報サービスとしてWWWを提供するサーバであり、ＨＴＭＬファイルをＵＲＬで関連付けて保存しており、また、画像ファイルや音声ファイルも保存している。WWWサーバ１はインターネット回線網２に接続されている。

【０００９】インターネット回線網２にはＣＴＩ(Computer Telephony Integration)サーバ３が接続されている。ＣＴＩサーバ３は公衆電話回線網４にも接続されている。公衆電話回線網４には複数の電話機が実際には接続されているが、ここでは１つの電話機５を示している。電話機は一般加入電話機、公衆電話機及び携帯電話機のいずれであっても良い。なお、公衆電話回線網４には中継局、基地局等の電話回線接続のための局が存在するが、図には示していない。

【００１０】ＣＴＩサーバ３はＷｅｂページの読み上げを電話機５を含む電話機のユーザに提供するサーバである。ＣＴＩサーバ３には、Ｗｅｂページ取得部３１と、テキスト抽出部３２と、テキスト読み上げ部３３とが備えられている。Ｗｅｂページ取得部３１はWWWサーバ１にアクセスし、Ｗｅｂページのソースデータを取得する。テキスト抽出部３２はＷｅｂページ取得部３１によって取得されたソースデータを解析し、テキスト部分を抽出する。テキスト読み上げ部３３はテキストの文字コードに応じて合成音声信号を作成し、その合成音声信号を公衆電話回線網４を利用して電話機に対して出力する。Ｗｅｂページ取得部３１、テキスト抽出部３２及びテキスト読み上げ部３３はＣＴＩサーバ３のプロセッサ（図示せず）の後述の如き動作によって形成される。

【００１１】また、ＣＴＩサーバ３は内部にハードディスク等の記憶装置３５を有しており、後述するように、ソースデータ等の各種データが記憶装置３５には記憶される。WWWサーバ１及びＣＴＩサーバ３各々のインターネット回線網２を利用した通信においては通信プロトコルとしてＴＣＰ／ＩＰが用いられ、WWWサーバ１及びＣＴＩサーバ３にはＩＰアドレスが各々割り当てられている。更に、WWWのプロトコルとしてはＨＴＴＰが使用される。また、図示してないが、WWWサーバ１及びＣＴＩサーバ３はルータを介してインターネット回線網２には接続されている。

【００１２】次に、かかるＷｅｂページ解読システムの動作について説明する。ユーザが電話機５からＣＴＩサーバ３へ電話をかけ、電話機５とＣＴＩサーバ３との間の通話状態が確立すると、ＣＴＩサーバ３は図２に示すように、先ず、予め定められたＵＲＬで指定される領域のＷｅｂページのソースデータを取得するために、そのＵＲＬに対してアクセスを行う（ステップＳ１）。このＵＲＬがWWWサーバ１内にあるとすると、WWWサーバ１はＵＲＬで指定される領域のＨＴＭＬファイル等のファイルからなるソースデータを読み出してＣＴＩサーバ３に対して送信する。そのソースデータはＷｅｂページを構成する基本となるソースデータである。送信されたソースデータはインターネット回線網２を介してＣＴＩサーバ３に供給される。

【００１３】ＣＴＩサーバ３はWWWサーバ１から送られて来たソースデータを記憶装置３５に格納し（ステップＳ２）、その格納したソースデータ中からＦＲＡＭＥ（フレーム）タグを抽出する（ステップＳ３）。このＦＲＡＭＥタグの抽出動作については後述するが、ＦＲＡＭＥタグ中に含まれるＵＲＬがフレームリストとして記憶装置３５に書き込まれる。

【００１４】ＣＴＩサーバ３はステップＳ３の実行の結果として、ＦＲＡＭＥタグの抽出が行われたか否かを判別する（ステップＳ４）。ステップＳ４にてＦＲＡＭＥタグの抽出が実際に行われなかった場合には、後述のステップＳ７に進む。一方、ＦＲＡＭＥタグの抽出が実際に行われた場合には、フレームリストのＵＲＬで指定される領域のＷｅｂページのソースデータを取得するために、そのフレームリストのＵＲＬに対してアクセスを行い（ステップＳ５）、WWWサーバ１から送られて来たソースデータを記憶装置３５に格納する（ステップＳ６）。ステップＳ５のアクセスに対するWWWサーバ１の動作はステップＳ１のアクセスの場合と同様である。ステップＳ６の実行後はステップＳ７に進む。

【００１５】ＣＴＩサーバ３はステップＳ７において記

憶装置３５に格納されたソースデータからタブ＜＞で囲まれた部分以外のテキスト部分を抽出し、その抽出テキストデータを読み上げデータとして記憶装置３５に書き込む（ステップＳ８）。その後、読み上げデータに基づいて合成音声信号を作成し（ステップＳ９）、その合成音声信号を電話機５に対して出力する（ステップＳ１０）。記憶装置３５に書き込まれた読み上げデータは複数の文字コードからなるテキストデータであるので、その文字コード各々又は単語単位の文字コード群に対応する音声データを記憶装置３５から検索して得て、それら音声データを合成して連続する合成音声信号を作成する。合成音声信号は公衆電話回線網４を介して電話機５に供給され、電話機５の受話器から読み上げ音が出力される。なお、記憶装置３５には文字コードと音声データとの関係を示すデータテーブルが予め記憶されている。

【００１６】次に、上記のステップＳ３におけるソースデータ中からのＦＲＡＭＥタグ抽出動作について図３のフローチャートを参照しつつ説明する。ＣＴＩサーバ３は、記憶装置３５に記憶されたソースデータ中から文字列＜ＦＲＡＭＥ　ＳＲＣを検索し（ステップＳ１１）、文字列＜ＦＲＡＭＥ　ＳＲＣがソースデータ中に存在するか否かを判別する（ステップＳ１２）。すなわち、記憶装置３５に書き込まれたソースデータ中にはＨＴＭＬファイルが含まれ、そのＨＴＭＬファイルが示す文書でフレーム設定が行われているか否かが判別される。文字列＜ＦＲＡＭＥ　ＳＲＣが存在するならば、＜ＦＲＡＭＥ　ＳＲＣ＝”ＵＲＬ”＞の構文が存在するので、読み取り位置をその次の文字＝の位置まで移動し（ステップＳ１３）、更に、その後の”　”で囲まれた文字列、すなわちＵＲＬをソースデータから読み取り、そのＵＲＬを記憶装置３５に形成されたフレームリストに書き込む（ステップＳ１４）。よって、フレームリストにはフレーム内に含まれるＨＴＭＬ文書の存在位置を示すＵＲＬが書き込まれる。ステップＳ１４の実行後、ソースデータの全てのファイルから文字列＜ＦＲＡＭＥ　ＳＲＣの検索が終了したか否かを判別し（ステップＳ１５）、その検索が終了していない場合にはステップＳ１１に戻り、上記のステップ動作を繰り返す。

【００１７】次いで、上記したステップＳ７におけるテキストデータの抽出動作について図４のフローチャートを参照しつつ説明する。ＣＴＩサーバ３は、記憶装置３５に格納されたソースデータのうちの１つのファイルの先頭から順に１文字分の文字コードを取得し（ステップＳ２１）、その文字コードが文字＜を示すか否かを判別する（ステップＳ２２）。取得した文字コードが文字＜を示す場合にはタグフラグF_{TAG}を１に等しくさせる（ステップＳ２３）。取得した文字コードが文字＜を示さない場合にはタグフラグF_{TAG}が１に等しいか否かを判別する（ステップＳ２４）。タグフラグF_{TAG}はＨＴＭＬファイルにおいて＜＞で囲まれた部分において１に設定され、それ以外の部分において０に設定されるフラグであり、その初期値は０である。ステップＳ２３の実行後もステップＳ２４の判別は実行される。

【００１８】ステップＳ２４の判別の結果、タグフラグF_{TAG}が１に等しくされている場合には、ステップＳ２１で取得した文字コードが文字＞を示すか否かを判別する（ステップＳ２５）。取得した文字コードが文字＞を示す場合にはタグの終了であるので、タグフラグF_{TAG}を０に等しくさせる（ステップＳ２６）。一方、ステップＳ２４の判別の結果、タグフラグF_{TAG}が０に等しくされている場合には、＜＞で囲まれたタグ部分以外のテキスト部分であるので、取得した文字コードを読み上げデータに含ませるように記憶装置３５に格納する（ステップＳ２７）。

【００１９】ステップＳ２６又はＳ２７の実行後はソースデータの全てから１文字分の文字コードの取得が終了したか否かを判別し（ステップＳ２８）、その取得が終了していない場合にはステップＳ２１に戻り、上記のステップ動作を繰り返す。よって、かかるシステムによれば、基本のソースデータがＦＲＡＭＥタグを含むＨＴＭＬファイルの場合にＦＲＡＭＥタグ内に記述されたＵＲＬの領域に格納されているＨＴＭＬファイルのテキスト部分も抽出することができるので、ＷＷＷブラウザに表示されるＷｅｂページのテキスト部分を余すことなく読み上げることができる。

【００２０】上記した実施例においては、フレームタグが使用された場合について説明したが、フレームタグ以外のＵＲＬの記述を有するタグにも本発明を適用することができる。また、ＪａｖａＳｃｒｉｐｔ等のスクリプト言語を含むＨＴＭＬファイルからテキスト部分を抽出して読み上げる場合にも本発明を適用することができる。ＨＴＭＬでは、機能を拡張するためにＷｅｂページ上でＪａｖａＳｃｒｉｐｔ等のスクリプト言語を実行できるようにするタグも用意されている。例えば、＜ＳＣＲＩＰＴ　ＬＡＮＧＵＡＧＥ＝”ＪａｖａＳｃｒｉｐｔ”＞～＜／ＳＣＲＩＰＴ＞のような構文で形成される。よって、上記したように＜ＳＣＲＩＰＴ　ＬＡＮＧＵＡＧＥ＝”ＪａｖａＳｃｒｉｐｔ”＞～＜／ＳＣＲＩＰＴ＞の範囲の部分を無視してそれ以外のテキスト部分を抽出するのである。

【００２１】更に、本発明のシステムはＨＴＭＬを用いたファイルに場合に限らず、ＨＴＭＬを拡張させた言語を用いたファイルについても適用することができる。なお、日本ではＷＷＷブラウザにて閲覧できるページを全てホームページと称しているが、ホームページは本来、１情報群を構成する複数のＷｅｂページのうちの基本ページであるので、ここでは誤解を招かないようにＷｅｂページと記載した。

【００２２】

【発明の効果】以上の如く、本発明のＷｅｂページ解読

システムにおいては、ＵＲＬの記述を有するタグを含むＨＴＭＬの文書に対してテキスト部分を正確に抽出することができる。よって、Ｗｅｂページを読み上げる際にはＷｅｂページのテキスト部分を余すことなく読み上げることができる。

【図面の簡単な説明】

【図１】本発明によるＷｅｂページ解読システムの構成を示すブロック図である。

【図２】図１のシステム中のＣＴＩサーバの動作を示すフローチャートである。

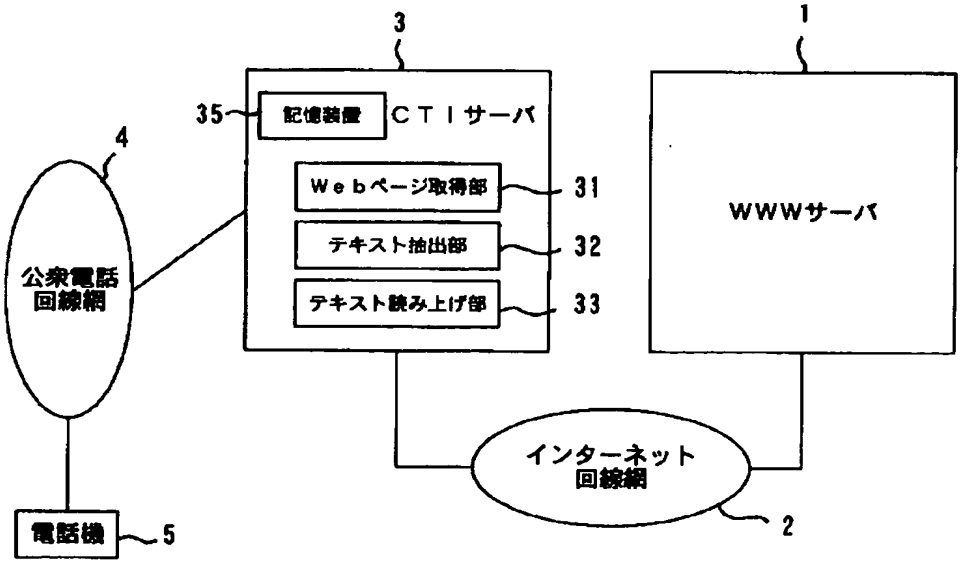
＊【図３】ソースデータ中からのＦＲＡＭＥタグ抽出動作を示すフローチャートである。

【図４】テキストデータの抽出動作を示すフローチャートである。

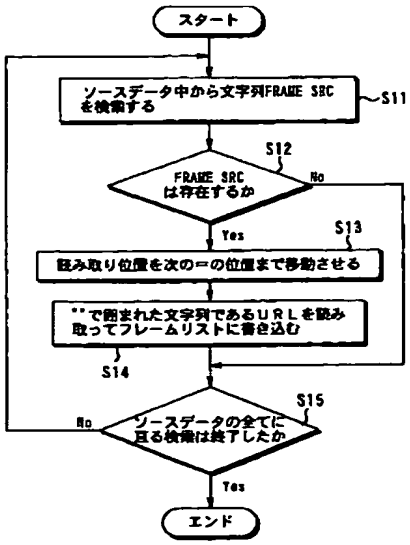
【符号の説明】

1　ＷＷＷサーバ
2　インターネット回線網
3　ＣＴＩサーバ
4　公衆電話回線網
5　電話機

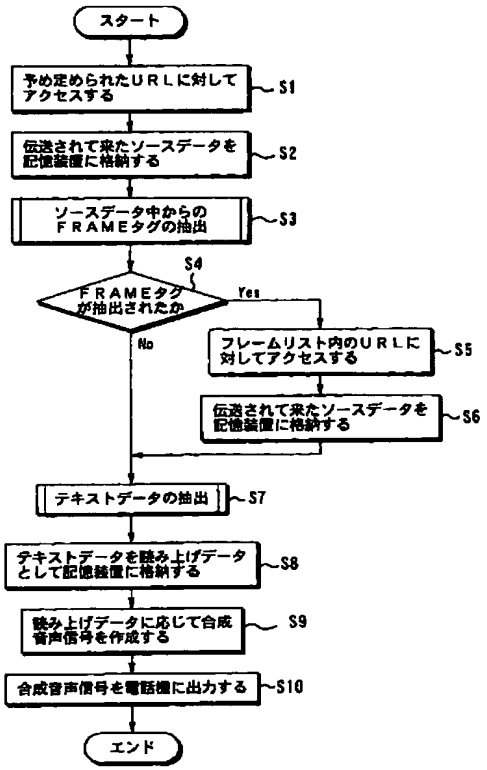
【図１】



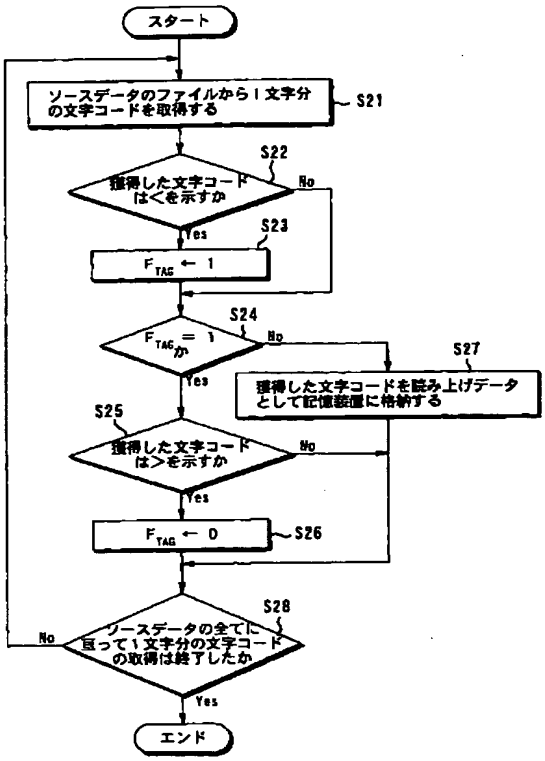
【図３】



【図２】



【図４】



フロントページの続き

(72)発明者　西村　仁司
東京都板橋区舟渡１丁目12番11号　沖ソフトウェア株式会社内
(72)発明者　八木田　一彦
東京都板橋区舟渡１丁目12番11号　沖ソフトウェア株式会社内
(72)発明者　松下　有亨
東京都板橋区舟渡１丁目12番11号　沖ソフトウェア株式会社内
(72)発明者　山口　雄一郎
東京都港区虎ノ門１丁目７番12号　沖電気工業株式会社内
(72)発明者　伊藤　慎一
東京都港区虎ノ門１丁目７番12号　沖電気工業株式会社内
(72)発明者　永井　友康
東京都千代田区大手町２－２－２　アーバンネット大手町ビル　株式会社エヌ・ティ・ティエムイー内
(72)発明者　大田　剛志
東京都千代田区大手町２－２－２　アーバンネット大手町ビル　株式会社エヌ・ティ・ティエムイー内
(72)発明者　田村　賢
東京都千代田区大手町２－２－２　アーバンネット大手町ビル　株式会社エヌ・ティ・ティエムイー内

Ｆターム(参考)　5B009 QA11 RD03 SA03 SA14 TA11
VA02 VC01
5B089 GA11 GB03 HA01 JA22 JB02
KA04 KB07 KC53 KC59 LB13

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-273214

(43)Date of publication of application : 05.10.2001

(51)Int.Cl. G06F 13/00
G06F 17/21

(21)Application number : 2000-083150

(22)Date of filing : 24.03.2000

(71)Applicant : OKI SOFTWARE KK
OKI ELECTRIC IND CO LTD
NTT ME CORP

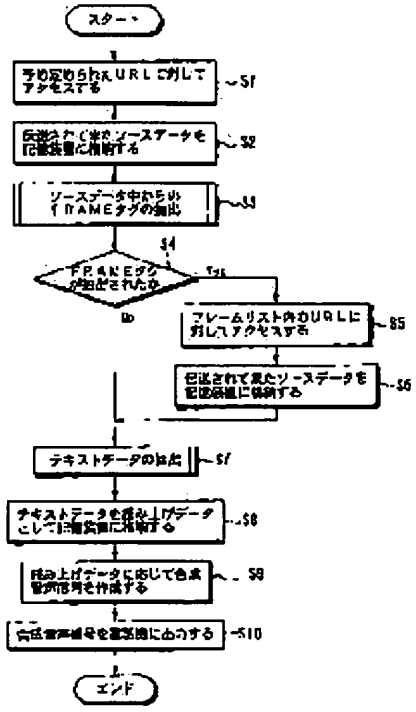
(72)Inventor : NISHIMURA HITOSHI
YAKIDA KAZUHIKO
MATSUSHITA ARIYUKI
YAMAGUCHI YUICHIRO
ITO SHINICHI
NAGAI TOMOYASU
OTA TSUYOSHI
TAMURA MASARU

(54) WEB PAGE DECODING SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide a Web page decoding system capable of exactly extracting a text part in respect to the document of an HTML containing a tag having the description of a uniform resource locator (URL).

SOLUTION: Basic source data comprising a Web page are extracted from a storage area designated by the prescribed URL, and written in a storage means and when the existence of the tag containing the description of the URL is detected out of the basic source data, the URL in that prescribed tag is detected. Then, source data are extracted from the storage area designated by that detected URL, and written in the storage means and the text part is extracted from all the source data stored



in the storage means.

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

CLAIMS

[Claim(s)]
[Claim 1]A Web page decipherment system which decodes a text part of an HTML document which constitutes a Web (web) page, comprising:
A means which takes out source data of foundations which constitute said Web page from a storage area specified by predetermined URL (uniform resource locator), and is written in a memory measure.
A tag detection means which detects existence of a predetermined tag including a description part of URL out of source data of said foundations.
A URL detection means to detect URL in the predetermined tag when existence of said predetermined tag is detected.
A means which takes out source data from a storage area specified by URL detected by said URL detection means, and is written in said memory measure, and a sampling-of-text means to extract a text part from all the source data memorized by said memory measure.

[Claim 2]The web page decipherment system according to claim 1, wherein said predetermined tag is a frame tag.
[Claim 3]The web page decipherment system according to claim 1 or 2, wherein said URL detection means detects URL in syntax of <FRAMESRC="URL"> as URL in said frame tag.
[Claim 4]The web page decipherment system according to claim 1, wherein said sampling-of-text means extracts portions other than a portion surrounded by <> in source data memorized by said memory measure as a text part.
[Claim 5]The web page decipherment system according to claim 1 having further a voice output means which creates and outputs an audio signal corresponding to a text part extracted by said sampling-of-text means.
[Claim 6]The web page decipherment system according to claim 1, wherein an output sound

signal of said voice output means is supplied to telephone via a dial-up line.

[Translation done.]

* NOTICES *

1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DETAILED DESCRIPTION

[Detailed Description of the Invention]
[0001]
[The technical field to which an invention belongs] This invention relates to the Web page decipherment system which decodes the text part of a Web (web) page.
[0002]
[Description of the Prior Art]WWW (World Wide Web) which is one of the information services of the Internet, Using the HTML file described in a language called HTML (Hyper Text Markup Language), and URL (Uniform Resource Locator) which is the identifiers of the preserving position of the file, via the Internet A character, an image, Multimedia information, such as a sound, can be referred to. It is a Web page which is formed on a display screen by processing an HTML file by the inspection software called a WWW browser. The WWW server associates and saves the HTML file by URL, and the side which provides the information on WWW operates according to server application. The source data which contain an HTML file by the side provided with information, i.e., a client, (client computer) from desired URL using a WWW browser. The Web page which accessed (for example, a graphics file and a voice file) via the Internet, and was formed by the file in sauce can be referred to on a display screen.
[0003]The document of HTML is usually formed from the text part which consists of text documents, and the display information bureau formed with a tag. A tag is a sign which makes <> a couple and the syntax of HTML is formed like <HTML>- </HTML> using a tag. Various display information, including the size of the character of the text part displayed on a Web page, the kind of font, its character color, the background color of a Web page, an image file name, an image position, etc., is shown in the portion surrounded by tag <>.
[0004]Thus, since the portion surrounded by tag <> in the document of HTML is the control information for displaying not the portion displayed but a document, if the portion surrounded by tag <> is removed, it is usually common to become a mere text document. The system

which, on the other hand, reads out the document of the Web page displayed on a WWW browser may be formed on the Internet. This is because the document of a Web page is read out and it sends out to telephone as an audio signal, when the terminal of a client is not a computer but the telephone connected to the public line, for example. Extracting the text data portion excluding the portion surrounded by tag <> out of the HTML file for read-aloud as Web page read-aloud, compounding the voice data corresponding to the character code of the text data portion, and outputting as a series of audio signals is performed.

[0005]

[Problem(s) to be Solved by the Invention]There is <FRAMESET>- </FRAMESET> in HTML as syntax for forming a frame on a Web page, for example, and the divided screen is obtained in the Web page using this. From the HTML file the syntax was described to be, another HTML file is further called for every split screen, and a document is displayed. That is, URL (an HTML file name is included) is specified for every split screen with the tag like <FRAME solvent refined coal="URL">, and the contents of the HTML file which exists in the specified field of URL are displayed.

[0006]However, there was a problem that a text part could not be correctly extracted to the document of HTML containing the tag which has description of URL like such a frame tag before. Then, the purpose of this invention is to provide the Web page decipherment system which can extract a text part correctly to the document of HTML containing the tag which has description of URL.

[0007]

[Means for Solving the Problem]A Web page decipherment system of this invention is provided with the following.

A means which is a Web page decipherment system which decodes a text part of an HTML document which constitutes a Web page, takes out source data of foundations which constitute a Web page from a storage area specified by predetermined URL (uniform resource locator), and is written in a memory measure.

A tag detection means which detects existence of a predetermined tag including a description part of URL out of basic source data.

A URL detection means to detect URL in the predetermined tag when existence of a predetermined tag is detected, A means which takes out source data from a storage area specified by URL detected by a URL detection means, and is written in a memory measure, and a sampling-of-text means to extract a text part from all the source data memorized by memory measure.

In the case of an HTML file in which basic source data contain a frame tag, from this composition, a text part of an HTML file stored in a field of URL described in that frame tag can also be extracted.

[0008]
[Embodiment of the Invention]Hereafter, the example of this invention is described in detail, referring to drawings. Drawing 1 shows the composition of the Web page decipherment system by that of this invention. In this system, WWW server 1 is a server which provides WWW as an information service, the HTML file is associated and saved by URL, and the graphics file and the voice file are also saved. WWW server 1 is connected to the Internet line network 2.
[0009]The CTI (Computer Telephony Integration) server 3 is connected to the Internet line network 2. CTI server 3 is connected also to the dial-up line network 4. Although two or more telephones are actually connected to the dial-up line network 4, the one telephone 5 is shown here. Telephones may be any of an ordinary phone machine, a public telephone, and a portable telephone. Although the office for dialups, such as a relay station and a base station, exists in the dial-up line network 4, it is not shown in a figure.
[0010]CTI server 3 is a server which provides with read-aloud of a Web page the user of the telephone containing the telephone 5. CTI server 3 is equipped with the Web page acquisition part 31, the sampling-of-text part 32, and the text read-aloud part 33. The Web page acquisition part 31 accesses WWW server 1, and acquires the source data of a Web page. The sampling-of-text part 32 analyzes the source data acquired by the Web page acquisition part 31, and extracts a text part. The text read-aloud part 33 creates a synthesized speech signal according to the character code of a text, and outputs the synthesized speech signal to telephone using the dial-up line network 4. The Web page acquisition part 31, the sampling-of-text part 32, and the text read-aloud part 33 are formed by the operation like the after-mentioned of the processor (not shown) of CTI server 3.
[0011]CTI server 3 has the memory storage 35, such as a hard disk, inside, and various data, such as source data, is memorized by the memory storage 35 so that it may mention later. in communication using the Internet line network 2 of WWW server 1 and CTI server 3 each, TCP/IP is used as a communications protocol, and the IP address is respectively assigned to WWW server 1 and CTI server 3. HTTP is used as a protocol of WWW. Although not illustrated, WWW server 1 and CTI server 3 are connected to the Internet line network 2 via the router.
[0012]Next, operation of this Web page decipherment system is explained. If a user telephones CTI server 3 from the telephone 5 and the talk state between the telephone 5 and CTI server 3 is established, as shown in drawing 2, CTI server 3, First, in order to acquire the source data of the Web page of the field specified by URL defined beforehand, it accesses to the URL (Step S1). Supposing this URL is in WWW server 1, WWW server 1 will read the source data which consist of files, such as an HTML file etc. of the field specified by URL, and will transmit to CTI server 3. The source data are the source data used as the foundations which constitute a Web page. The transmitted source data are supplied to CTI server 3 via the

Internet line network 2.

[0013]CTI server 3 stores in the memory storage 35 the source data sent from WWW server 1 (Step S2), and extracts a FRAME (frame) tag out of the stored source data (Step S3). Although the extraction operation of this FRAME tag is mentioned later, URL contained in a FRAME tag is written in the memory storage 35 as a frame list.

[0014]CTI server 3 distinguishes whether extraction of the FRAME tag was performed as a result of execution of Step S3 (step S4). When extraction of a FRAME tag is not actually performed in step S4, it progresses to the below-mentioned step S7. On the other hand, when extraction of a FRAME tag is actually performed, In order to acquire the source data of the Web page of the field specified by URL of a frame list, it accesses to URL of the frame list (Step S5), and the source data sent from WWW server 1 are stored in the memory storage 35 (Step S6). Operation of WWW server 1 to access of Step S5 is the same as that of the case of access of Step S1. After execution of Step S6 progresses to Step S7.

[0015]CTI server 3 extracts text parts other than the portion surrounded by tab <> from the source data stored in the memory storage 35 in Step S7, reads out the extraction text data, and writes it in the memory storage 35 as data (Step S8). Then, a synthesized speech signal is created based on read-aloud data (step S9), and the synthesized speech signal is outputted to the telephone 5 (Step S10). since the read-aloud data written in the memory storage 35 is text data which consists of two or more character codes, the voice data corresponding to the character code group of the character codes of each and word unit is searched and obtained from the memory storage 35, and the synthesized speech signal which compounds these voice data and continues is created. A synthesized speech signal is supplied to the telephone 5 via the dial-up line network 4, it reads out from the receiver of the telephone 5, and a sound is outputted. The data table showing the relation between a character code and voice data in the memory storage 35 is memorized beforehand.

[0016]Next, it explains, referring to the flow chart of drawing 3 for the FRAME tag extraction operation out of the source data in the above-mentioned step S3. CTI server 3 searches character string <FRAME solvent refined coal out of the source data memorized by the memory storage 35 (Step S11), and distinguishes whether character string <FRAME solvent refined coal exists in source data (Step S12). That is, an HTML file is contained in the source data written in the memory storage 35, and it is distinguished whether frame settings are performed by the document which the HTML file shows. Since the syntax of <FRAME solvent refined coal="URL"> exists if character string <FRAME solvent refined coal exists, A read position is moved to the position of the following character = (Step S13), the character string surrounded by subsequent "", i.e., URL, is further read in source data, and the URL is written in the frame list formed in the memory storage 35 (Step S14). Therefore, URL which shows the existence position of the HTML document contained in a frame is written in a frame list. It

distinguishes whether search of character string <FRAME solvent refined coal was completed from all the files of source data after execution of Step S14 (Step S15), when the search is not completed, it returns to Step S11, and the above-mentioned step operation is repeated.

[0017]Subsequently, it explains, referring to the flow chart of drawing 4 for the extraction operation of the text data in the above-mentioned step S7. CTI server 3 acquires the character code for one character sequentially from the head of one file in the source data stored in the memory storage 35 (Step S21) -- the character code -- a character -- < -- it is distinguished whether it is shown or not (Step S22). the acquired character code -- a character -- < -- when shown, tag flag F_{TAG} is made equal to one (Step S23). the acquired character code -- a character -- < -- when not shown, it is distinguished whether tag flag F_{TAG} is [one] equal (Step S24). It is a flag which tag flag F_{TAG} is set as 1 in the portion surrounded by <> in the HTML file, and is set as 0 in the other portion, and the initial value is 0. Distinction of Step S24 is performed even after execution of Step S23.

[0018]As a result of distinction of Step S24, when tag flag F_{TAG} is made equal to one, the character code acquired at Step S21 distinguishes whether character > is shown (Step S25). Since it is the end of a tag when the acquired character code shows character >, tag flag F_{TAG} is made equal to zero (Step S26). On the other hand, as a result of distinction of Step S24, since it is text parts other than the tag portion surrounded by <> when tag flag F_{TAG} is made equal to zero, the acquired character code is read out, and it stores in the memory storage 35 so that it may be made to contain in data (Step S27).

[0019]After Step S26 or execution of S27 distinguishes whether acquisition of the character code for one character was completed from all the source data (Step S28), when the acquisition is not completed, it returns to Step S21, and it repeats the above-mentioned step operation. Therefore, since the text part of the HTML file stored in the field of URL which was described in the FRAME tag in the case of the HTML file in which basic source data contain a FRAME tag can also be extracted according to this system, It can read out without leaving the text part of the Web page displayed on a WWW browser.

[0020]In the above-mentioned example, although the case where a frame tag was used was explained, this invention is applicable also to the tag which has description of URL other than a frame tag. This invention can be applied also when extracting and reading out a text part from the HTML file containing script languages, such as JavaScript. In HTML, in order to extend a function, the tag which enables it to perform script languages, such as JavaScript, on a Web page is also prepared. For example, it is formed by syntax like <SCRIPT LANGUAGE="JavaScript">- </SCRIPT>. Therefore, as described above, the portion of the range of <SCRIPT LANGUAGE="JavaScript">- </SCRIPT> is disregarded and the other text

part is extracted.

[0021]The system of this invention is applicable also to the file using the language to which the file which used HTML was made to extend not only a case but HTML. Although all the pages that can be perused in a WWW browser were called the homepage in Japan, since the homepage was originally a basic page of two or more Web pages which constitute one information group, it was indicated with the Web page that misunderstanding was not invited here.

[0022]

[Effect of the Invention]Like the above, a text part can be correctly extracted in the Web page decipherment system of this invention to the document of HTML containing the tag which has description of URL. Therefore, it can read out, without leaving the text part of a Web page, when reading out a Web page.

[Translation done.]

DESCRIPTION OF DRAWINGS

[Brief Description of the Drawings]
[Drawing 1]It is a block diagram showing the composition of the Web page decipherment system by this invention.
[Drawing 2]It is a flow chart which shows operation of the CTI server in the system of drawing 1.
[Drawing 3]It is a flow chart which shows the FRAME tag extraction operation out of source data.
[Drawing 4]It is a flow chart which shows the extraction operation of text data.
[Description of Notations]
1 WWW server
2 Internet line network
3 CTI server
4 Dial-up line network
5 Telephone

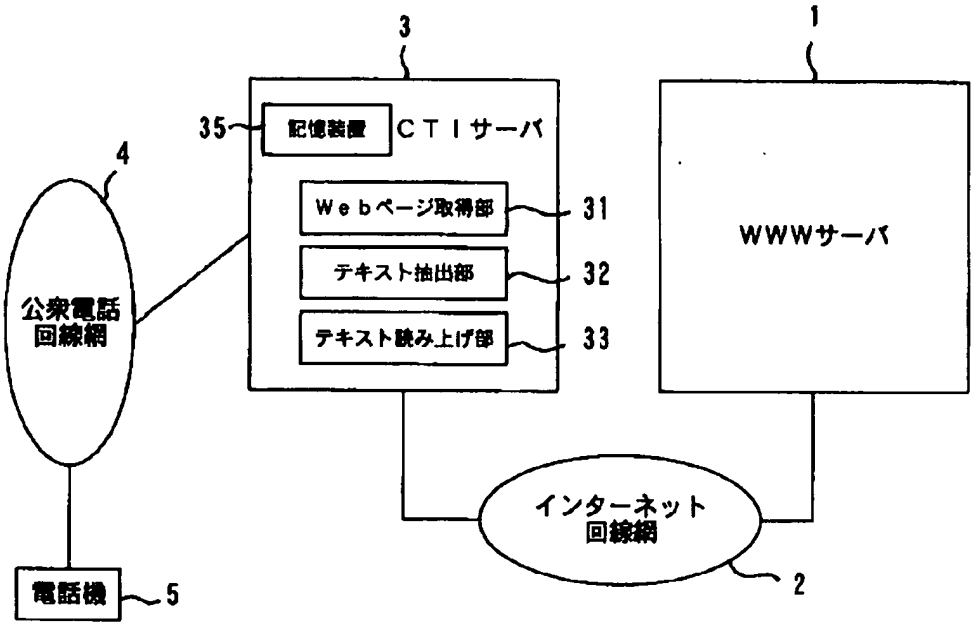
[Translation done.]

* NOTICES *

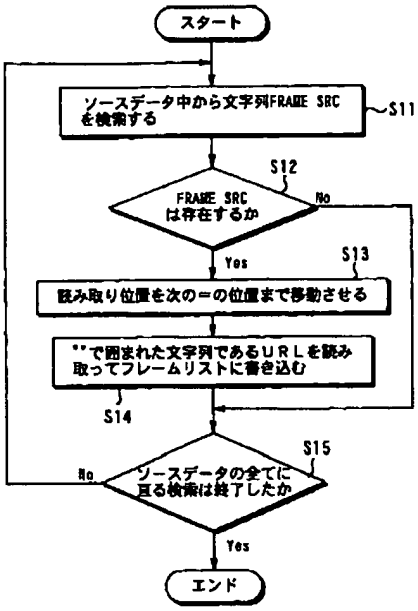
1.This document has been translated by computer. So the translation may not reflect the original precisely.
2.**** shows the word which can not be translated.
3.In the drawings, any words are not translated.

DRAWINGS

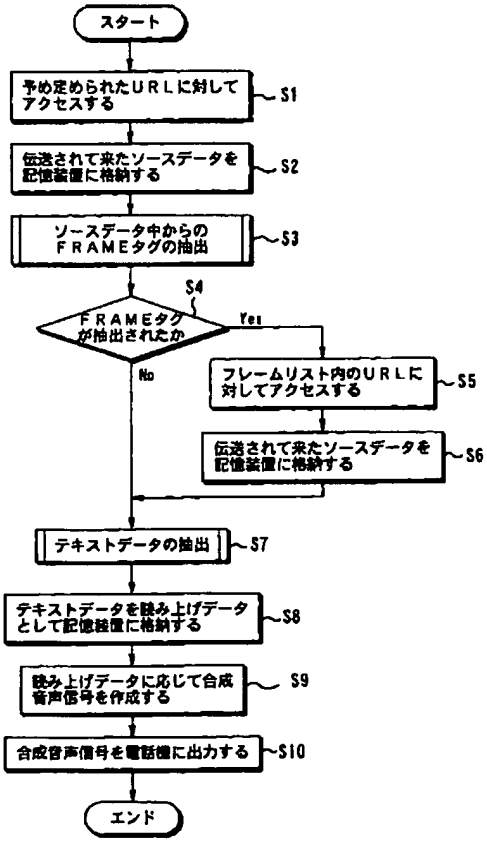
[Drawing 1]



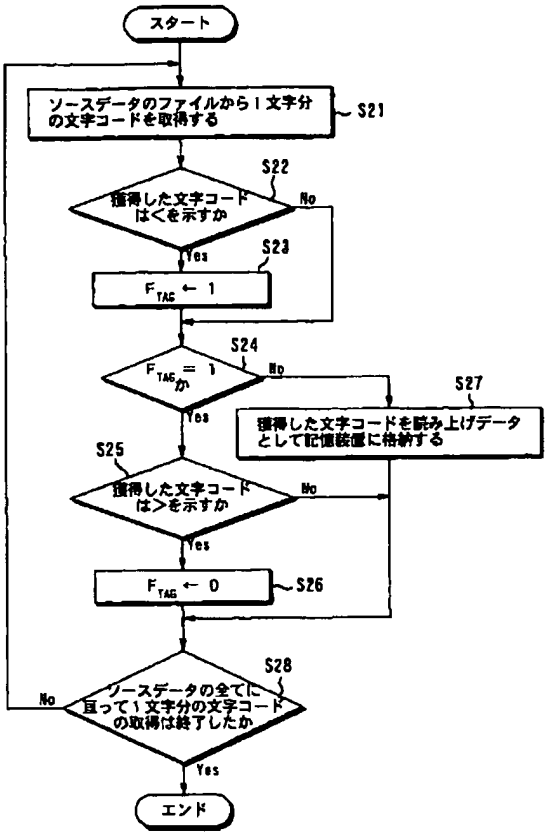
[Drawing 3]



[Drawing 2]



[Drawing 4]



[Translation done.]